

# Bayesian inference with monotone instrumental variables

Hang Qian

Iowa State University

August 2011

# Bayesian Inference with Monotone Instrumental Variables

Hang Qian

**Abstract**

Sampling variations complicate the classical inference on the analogue bounds under the monotone instrumental variables assumption, since point estimators are biased and confidence intervals are difficult to construct. From the Bayesian perspective, a solution is offered in this paper. Using a conjugate Dirichlet prior, we derive some analytic results on the posterior distribution of the two bounds of the conditional mean response. The bounds of the unconditional mean response and the average treatment effect can be obtained with Bayesian simulation techniques. Our Bayesian inference is applied to an empirical problem which quantifies the effects of taking extra classes on high school students' test scores. The two MIVs are chosen as the education levels of their fathers and mothers. The empirical results suggest that the MIV assumption in conjunction with the monotone treatment response assumption yield good identification power.

*Keywords:* Monotone instrumental variables, Bayesian, Dirichlet.

## 1. Introduction

Identification of treatment effects requires assumptions imposed on the joint distribution of treatment and response variables as well as covariates.

Under the monotone instrumental variable (MIV) assumption introduced by Manski and Pepper (2000), mean responses vary monotonically across specified sub-populations defined by the MIV. It has a wide application potential since an MIV is less restrictive and easier to provide than a valid instrumental variable. However, it comes across a barrier in applications. One feature of the MIV model is that a supremum (infimum) operator will appear in the sharp lower (upper) bound. As is noted by Manski and Pepper (2009, p.211), "the sup and inf operations ... significantly complicate the bounds under other MIV assumptions, rendering it difficult to analyze the sampling behavior of analogue estimates. Moreover, the methods for forming asymptotically valid confidence sets for partially identified parameters ... appear not to apply." [1] In a frequentist framework, the true bounds, fixed but unknown, are functions of population moments or probabilities. Problems arise when we use the sample analogues to replace the true bounds, since Jensen's inequality indicates that the analogue estimate of the lower (upper) bound will be biased upwards (downwards), resulting in the estimates narrower than the true bounds. To resolve this complication, Kreider and Pepper (2007) propose a heuristic bootstrap bias correction, and Qian (2010) provides a justification for that approach and extend it to multi-level simultaneous bootstraps. Chernozhukov et al. (2009) develop an inference method on intersection bounds with a continuum of inequalities. Their estimator maximizes or minimizes the precision-corrected curve defined by the

[1] The bounds under the monontone treatment selection assumption have simple forms, but under other MIV assumptions the supremum and infimum operators will appear in the bounds.

analogue estimates plus a critical value multiplied by pointwise standard errors.

In this paper, a Bayesian solution is offered. We argue that the complication is not induced by the sampling variation, but by the way we interpret our uncertainty on the bounds. Bayesians interpret the probability as a degree of belief, and therefore the MIV bounds themselves are random. It is likely that little is known about the bounds prior to the data being observed. Learning from the data, we sharpen our understandings on the MIV bounds. In that sense, our posterior belief on the MIV bounds are simply the supremum or infimum of a set of random variables. Once we find out their posterior distributions, we can articulate, for example, the most likely value of the MIV upper bound, the interval which we are 95% sure the upper bound will fall into. Since our knowledge on the bounds updates with the data, there never exists an absolutely true bound, and therefore there is no biased or unbiased belief.

The main contribution of this paper is to derive some finite sample analytic distributions of the MIV bounds. We begin by discretizing the treatment variable, response variable as well as the MIV. With a conjugate Dirichlet prior, we arrive at posterior probabilities. We then show that the MIV bounds are the maximum or minimum of a set of random variables, each of which is a linear combination of the conditional probabilities. With the Gamma representation of the Dirichlet posteriors, it is possible to find out the closed-form distribution function. Our work is closely related to a body of statistics literature studying linear combinations of distributions in common families. Imhof (1961) computes the distribution of quadratic form in normal

variates. Moschopoulos (1985) provides algorithms on computing the distribution function of linear combination of the gamma family. Provost and Cheong (2000) study the linear combination of components of a Dirichlet vector.

Discretization of all variables is arguable. First, the treatment variable is discrete, usually binary, in most applications. Second, the MIV identification requires the response variable is bounded below and above. Otherwise the MIV has no identification power unless it is used together with monotone treatment selection (Manski, 1997). Lower and upper bounds are readily available when the response variable is discretized into grids. Third, to compute the analogue estimates for each subpopulation classified by the MIV, we usually group the values of the MIV so as to ensure enough sample size, which results in a discretized MIV.

The rest of the paper is organized as follows. Section 2 reviews the structure of the MIV bounds in Manski and Pepper (2000). Section 3 describes a Bayesian inference on the distribution of bounds under the MIV assumption. Section 4 extends the Bayesian inference method to other identification problems such as those in Kreider and Pepper (2007). Section 5 provides an application about bounding the impact of two factors on the test scores of high school students. Posterior bounds on the effects of taking extra classes, as well as the education level of parents are identified.

## 2. The structure of the MIV bounds

A discrete version of the counterfactual outcomes identification problem in Manski and Pepper (2000) can be raised as follows. Let $D \in \{d_1, ..., d_{n_D}\}$

be a treatment variable. For each treatment variety, there is a corresponding latent response variable denoted as $Y_t \in \{y_1, ..., y_{n_Y}\}$, $t = 1, ..., n_D$. Since a person can receive only one variety of treatment, the only observable outcome is $Y = \sum_{t=1}^{n_D} Y_t \cdot I(D = d_t)$, where $I(\cdot)$ is an indicator function. Let $Z \in \{z_1, ...z_{n_Z}\}$ be a MIV such that for any two realizations $z_i \leq z_j$,

$$E(Y_t | Z = z_i) \leq E(Y_t | Z = z_j), \forall t = 1, ..., n_D.$$

Without loss of generality, the values of each variable are sorted in an increasing order: $d_1 \leq ... \leq d_{n_D}$, $y_1 \leq ... \leq y_{n_Y}$, $z_1 \leq ... \leq z_{n_Z}$.

To bound $E(Y_t | Z = z_j)$, for some $t = 1, ..., n_D$, $j = 1, ..., n_Z$, we immediately have

$$\sup_{1 \leq i \leq j} E(Y_t | Z = z_i) \leq E(Y_t | Z = z_j) \leq \inf_{j \leq i \leq n_Z} E(Y_t | Z = z_i).$$

Note that the MIV is discretized, below we will use $\max(\cdot)$, $\min(\cdot)$ instead of $\sup(\cdot)$, $\inf(\cdot)$. However, $E(Y_t | Z = z_i)$ cannot be directly estimated due to counterfactuals. We dissemble it into an observable part $E(Y | Z = z_i, D = d_t)$ and an unobservable part $E(Y_t | Z = z_i, D \neq d_t)$. The latter need to be replaced by the worse-case lower bound $y_1$ and upper bound $y_{n_Y}$, which yield the sharp bounds under the MIV assumption alone:

$$\max_{1 \leq i \leq j} E(Y | Z = z_i, D = d_t) \cdot P(D = d_t | Z = z_i) + y_1 \cdot P(D \neq d_t | Z = z_i) \tag{1}$$

$$\leq E(Y_t | Z = z_j) \leq$$

$$\min_{j \leq i \leq n_Z} E(Y | Z = z_i, D = d_t) \cdot P(D = d_t | Z = z_i) + y_{n_Y} \cdot P(D \neq d_t | Z = z_i).$$

Since variables have been discretized, we can expand the conditional expectation in terms of conditional probabilities. To make notations compact,

let us define

$p_{ikm} \equiv P\left(Z = z_i, Y = y_k, D = d_m\right)$,

$i = 1, ..., n_Z$, $k = 1, ..., n_Y$, $m = 1, ..., n_D$,

$p_{i\cdot\cdot} \equiv \sum_{k=1}^{n_Y} \sum_{m=1}^{n_D} p_{ikm}$,

$x_{ikm} \equiv \frac{p_{ikm}}{p_{i\cdot\cdot}}$, where we assume $p_{i\cdot\cdot} > 0$, $\forall i = 1, ..., n_Z$. So $x_{ikm}$ stands for the conditional probability $P\left(Y = y_k, D = d_m \,|Z = z_i\,\right)$,

$\mathbf{p} \equiv vec\left(\{p_{ikm}\}_{i=1,k=1,m=1}^{n_Z,n_Y,n_D}\right)$,

$\mathbf{p}_i \equiv vec\left(\{p_{ikm}\}_{k=1,m=1}^{n_Y,n_D}\right)$.

Here we use the operator $vec\left(\cdot\right)$ to vectorize the multi-dimension array into a long column vector. For instance, $vec\left(\{p_{ikm}\}_{i=1,k=1,m=1}^{n_Z,n_Y,n_D}\right)$ turns a $n_Z \times n_Y \times n_D$ array to a vector of length $n_Z n_Y n_D$.

Then Eq. (1) can be written as

$$\max_{1 \le i \le j} f_L\left(\mathbf{x}_i\right) \le E\left(Y_t \,|Z = z_j\,\right) \le \min_{j \le i \le n_Z} f_U\left(\mathbf{x}_i\right), \tag{2}$$

where

$$\begin{aligned} f_L\left(\mathbf{x}_i\right) &= \sum_{k=1}^{n_Y} \sum_{m=1}^{n_D} \beta_{ikm} \cdot x_{ikm}, \\ f_U\left(\mathbf{x}_i\right) &= \sum_{k=1}^{n_Y} \sum_{m=1}^{n_D} \widetilde{\beta}_{ikm} \cdot x_{ikm}, \\ \beta_{ikm} &= y_k \cdot I\left(m = t\right) + y_1 \cdot I\left(m \neq t\right), \\ \widetilde{\beta}_{ikm} &= y_k \cdot I\left(m = t\right) + y_{n_Y} \cdot I\left(m \neq t\right). \end{aligned}$$

In words, the lower (upper) bound of $E\left(Y_t \,|Z = z_j\,\right)$ is the maximum (minimum) of a set of variables, each of which is a linear combination of conditional probabilities $x_{ikm}$ with combination coefficients either $y_k$ or $y_1$.

## 3. Bayesian inference

Bayesians treat $\mathbf{p}$ as a random vector. Before data are observed, our uncertainty over $\mathbf{p}$ can be modeled as a Dirichlet prior $Dir\left[vec\left(\{b_{ikm}\}_{i=1,k=1,m=1}^{n_Z,n_Y,n_D}\right)\right]$. If we lack prior information or show less subjectiveness on its prior distribution, we might choose each of the hyperparameters in $\{b_{ikm}\}_{i=1,k=1,m=1}^{n_Z,n_Y,n_D}$ to be 1 (uniform prior), or $\frac{1}{2}$ (Jeffreys' prior), or 0 (improper prior). Then we learn from the data, which are realizations from the multinomial distribution. Let $\{N_{ikm}\}_{i=1,k=1,m=1}^{n_Z,n_Y,n_D}$ be the number of occurrence of the type $(Z=z_i, Y=y_k, D=d_m)$ in the sample. It is well known that the posterior distribution of $\mathbf{p}$ is a conjugate $Dir\left[vec\left(\{a_{ikm}\}_{i=1,k=1,m=1}^{n_Z,n_Y,n_D}\right)\right]$, where $a_{ikm}=b_{ikm}+N_{ikm}$.

**Proposition 1.** *Posterior* $\mathbf{x}_i$, $i=1,...,n_Z$ *are independent Dirichlet vectors, and*

$$\mathbf{x}_i \sim Dir\left[vec\left(\{a_{ikm}\}_{k=1,m=1}^{n_Y,n_D}\right)\right].$$

Proof:

We use the Gamma representation of the Dirichlet distribution.

Let $\{\xi_{ikm}\}_{i=1,k=1,m=1}^{n_Z,n_Y,n_D}$ be an array of independently distributed random variables with each component $\xi_{ikm} \sim Gamma\left(a_{ikm},1\right)$.

Let $\eta=\sum_{i=1}^{n_Z}\sum_{k=1}^{n_Y}\sum_{m=1}^{n_D}\xi_{ikm}$, and $p_{ikm}=\frac{\xi_{ikm}}{\eta}$. By the change of variable method, we know $\mathbf{p}\equiv vec\left(\{p_{ikm}\}_{i=1,k=1,m=1}^{n_Z,n_Y,n_D}\right)$ has the Dirichlet distribution $Dir\left[vec\left(\{a_{ikm}\}_{i=1,k=1,m=1}^{n_Z,n_Y,n_D}\right)\right]$.

Consider the conditional probability

$$x_{ikm}=\frac{p_{ikm}}{p_{i\cdot\cdot}}=\frac{\xi_{ikm}}{\xi_{i\cdot\cdot}},$$

where $\xi_{i\cdot\cdot} = \sum_{k=1}^{n_Y} \sum_{m=1}^{n_D} \xi_{ikm}$.

Use the Gamma representation again, we recognize that for each $i = 1, ..., n_Z$, $\mathbf{x}_i \equiv vec\left(\{x_{ikm}\}_{k=1,m=1}^{n_Y,n_D}\right)$ has the Dirichlet distribution $Dir\left[vec\left(\{a_{ikm}\}_{k=1,m=1}^{n_Y,n_D}\right)\right]$. Also, $\mathbf{x}_1, ..., \mathbf{x}_{n_Z}$ use non-overlapping components in $\{\xi_{ikm}\}_{i=1,k=1,m=1}^{n_Z,n_Y,n_D}$, hence independence.■

**Proposition 2.** *The two bounds of $E(Y_t | Z = z_j)$ defined in Eq. (2), namely $L \equiv \max_{1\leq i\leq j} f_L(\mathbf{x}_i)$, $U \equiv \min_{j\leq i\leq n_Z} f_U(\mathbf{x}_i)$, has the posterior cumulative distribution function (c.d.f.)*

$$F_L(c) = \prod_{1\leq i\leq j} \left\{ \frac{1}{2} + \int_0^\infty \frac{1}{\pi s} \left[ \prod_{k=1}^{n_Y} \prod_{m=1}^{n_D} (r_{ikm})^{-a_{ikm}} \right] \cdot \sin\left( \sum_{k=1}^{n_Y} \sum_{m=1}^{n_D} a_{ikm}\theta_{ikm} \right) ds \right\},$$

$$F_U(c) = 1 - \prod_{1\leq i\leq j} \left\{ \frac{1}{2} - \int_0^\infty \frac{1}{\pi s} \left[ \prod_{k=1}^{n_Y} \prod_{m=1}^{n_D} (\widetilde{r}_{ikm})^{-a_{ikm}} \right] \cdot \sin\left( \sum_{k=1}^{n_Y} \sum_{m=1}^{n_D} a_{ikm}\widetilde{\theta}_{ikm} \right) ds \right\}$$

*where*

$$\begin{aligned}
r_{ikm} &= \sqrt{1 + (\beta_{ikm} - c)^2 s^2}, \\
\theta_{ikm} &= \arctan\left[ -(\beta_{ikm} - c) s \right], \\
\widetilde{r}_{ikm} &= \sqrt{1 + \left(\widetilde{\beta}_{ikm} - c\right)^2 s^2}, \\
\widetilde{\theta}_{ikm} &= \arctan\left[ -\left(\widetilde{\beta}_{ikm} - c\right) s \right].
\end{aligned}$$

Proof:

Let $L_i = f_L(\mathbf{x}_i)$, $i = 1, ...j$. It follows that $F_L(c) = \prod_{1\leq i\leq j} F_{L_i}(c)$.

Use the Gamma representation with $\{\xi_{ikm}\}_{i=1,k=1,m=1}^{n_Z,n_Y,n_D}$ defined in the proof of Proposition 1.

$$F_{L_i}(c) = P\left( \sum_{k=1}^{n_Y} \sum_{m=1}^{n_D} \beta_{ikm} \frac{\xi_{ikm}}{\xi_{i\cdot\cdot}} \leq c \right) = P(w_i \leq 0),$$

where

$$w_i = \sum_{k=1}^{n_Y} \sum_{m=1}^{n_D} (\beta_{ikm} - c)\, \xi_{ikm}.$$

To avoid confusion with the subscript $i$, denote the imaginary number $\iota = \sqrt{-1}$. Since $w_i$ is a linear combination of independent Gamma random variables, its characteristic function $\varphi_{w_i}(\cdot)$ takes the form

$$\begin{aligned}\varphi_{w_i}(s) &= \prod_{k=1}^{n_Y} \prod_{m=1}^{n_D} [1 - \iota \cdot (\beta_{ikm} - c)\, s]^{-a_{ikm}} \\ &= \prod_{k=1}^{n_Y} \prod_{m=1}^{n_D} (r_{ikm})^{-a_{ikm}} \cdot \exp\left(-\iota \cdot \sum_{k=1}^{n_Y} \sum_{m=1}^{n_D} a_{ikm}\theta_{ikm}\right),\end{aligned}$$

where $r_{ikm}, \theta_{ikm}$ are the polar representation of $1 - \iota \cdot (\beta_{ikm} - c)\, s$.

Then we use the inversion method proposed by Gil-Pelaez (1951),

$$\begin{aligned}P(w_i \leq 0) &= \frac{1}{2} - \int_0^\infty \frac{\mathrm{Im}\left[e^{-\iota s 0}\varphi_{w_i}(s)\right]}{\pi s} ds \\ &= \frac{1}{2} + \int_0^\infty \frac{1}{\pi s} \left[\prod_{k=1}^{n_Y} \prod_{m=1}^{n_D} (r_{ikm})^{-a_{ikm}}\right] \cdot \sin\left(\sum_{k=1}^{n_Y} \sum_{m=1}^{n_D} a_{ikm}\theta_{ikm}\right) ds.\end{aligned}$$

As for the upper bound, let $U_i = f_U(\mathbf{x}_i)$, $i = j, ..., n_Z$. It follows that $F_U(c) = 1 - \prod_{j \leq i \leq n_Z} [1 - F_{U_i}(c)]$, where $F_{U_i}(c)$ takes the same form as $F_{L_i}(c)$ with $\beta_{ikm}$ replaced by $\widetilde{\beta}_{ikm}$. ■

The integral above can be evaluated with deterministic Gaussian quadratures without difficulty. With the posterior c.d.f., it is straightforward to compute a 95% credible interval, especially the one with highest posterior density (HPD). Chen and Shao (1998) propose algorithms to find the Bayesian HPD regions. Frequentist confidence intervals for partially identified parameters are discussed in Imbens and Manski (2004), Chernozhukov

et al. (2007) and Rosen (2008). The advantage of the Bayesian interval is its simplicity. The interpretation is that the lower bound of $E\left(Y_m\left|Z=z_j\right.\right)$ falls into the credible interval with 95% probability, precisely in the finite sample.

If we are also interested in the posterior mean, we might use the formula suggested by David (1981) and Ross (2010).

$$E\left(L\right)=\int_0^\infty\left[1-F_L\left(c\right)-F_L\left(-c\right)\right]dc.$$

Note that the derived analytic distribution is for the bounds of conditional mean response, so Proposition 2 is most useful when the MIV defines a subpopulation of interest, especially when we concern two factors which may affect the potential outcomes–one is the treatment variable, the other is the MIV. In section 5 we give an application of this type. If we are interested in the bounds of unconditional mean response as well as the average treatment effect (ATE), we need to marginalize the conditional mean in accordance with the marginal distribution of the MIV. Analytic results are not available in that the marginal probabilities (i.e., $p_{j\cdot\cdot}, j=1,...,n_Z$) are also jointly Dirichlet random variables. When we take the product of $p_{j\cdot\cdot}$ and $\max_{1\leq i\leq j}f_L\left(\mathbf{x}_i\right)$, the resulting distribution is unknown. Despite this limitation, the Bayesian inference on the bounds of marginal mean response and the ATE can nevertheless be performed by simulation. We may start from the posterior distribution of $\mathbf{p}$. Random draws from $Dir\left[vec\left(\left\{a_{ikm}\right\}_{i=1,k=1,m=1}^{n_Z,n_Y,n_D}\right)\right]$ can be obtained by either the Gamma representation or consecutive draws from the marginal / conditional Beta distributions. For each draw of $\mathbf{p}$, we use Eq. (2) to obtain the simulated $\left\{f_L\left(\mathbf{x}_i\right)\right\}_{i=1}^{n_Z}$, $\left\{f_U\left(\mathbf{x}_i\right)\right\}_{i=1}^{n_Z}$. The same draw of $\mathbf{p}$ can also be used to obtain simulated $\left\{p_{j\cdot\cdot}\right\}_{j=1}^{n_Z}$. Then we compute $\sum_{j=1}^{n_Z}p_{j\cdot\cdot}\max_{1\leq i\leq j}f_L\left(\mathbf{x}_i\right)$ and $\sum_{j=1}^{n_Z}p_{j\cdot\cdot}\min_{j\leq i\leq n_Z}f_U\left(\mathbf{x}_i\right)$, which are simulated lower and upper bounds

of $E(Y_t)$. Repeating the process many times, we obtain i.i.d. draws from the posterior distribution of the two bounds of $E(Y_t)$. Therefore, we can find the posterior mean, median, HPD credible interval, etc. using those draws. Similarly, to simulate bounds of the ATE, say $E(Y_{t_1}) - E(Y_{t_2})$, we use one draw of $\mathbf{p}$ to compute the upper bound of $E(Y_{t_1})$ and the lower bound of $E(Y_{t_2})$ respectively. The difference is one draw from the posterior upper bound for the ATE. By repeated drawing, we learn its posterior distribution. The lower bound can be simulated similarly.

## 4. Extension

First, we provide two simple extensions, which will be used in our application in the next section. The MIV assumption is sometimes used together with the monotone treatment response (MTR) assumption (Manski, 1997) so that the identification power will be enhanced. The MTR implies, ceteris paribus, conjectured response varies monotonically with treatment everywhere in the sample space. With the MTR assumption, a better lower bound of $E(Y_t | Z = z_i, D = d_m)$, $m < t$ can be identified by $E(Y | Z = z_i, D = d_m)$ instead of the worst-case bound $y_1$. Similarly, to identify the upper bound of $E(Y_t | Z = z_i, D = d_m)$, $m > t$, we can use $E(Y | Z = z_i, D = d_m)$ instead of the worst-case bound $y_{n_Y}$.

As a result, the bounds of $E(Y_t | Z = z_j)$ under the MIV plus MTR as-

sumptions are

$$\max_{1\leq i\leq j}\sum_{k=1}^{n_Y}\sum_{m=1}^{n_D}x_{ikm}\left[y_k\cdot I\left(m\leq t\right)+y_1\cdot I\left(m>t\right)\right] \tag{3}$$
$$\leq E\left(Y_t\left|Z=z_j\right.\right)\leq$$
$$\min_{j\leq i\leq n_Z}\sum_{k=1}^{n_Y}\sum_{m=1}^{n_D}x_{ikm}\left[y_k\cdot I\left(m\geq t\right)+y_{n_Y}\cdot I\left(m<t\right)\right].$$

The Bayesian inference procedure is largely unchanged with the additional MTR assumption. Proposition 2 still applies, with the linear combination coefficients replaced by $\beta_{ikm}=y_k\cdot I\left(m\leq t\right)+y_1\cdot I\left(m>t\right)$ for the lower bound.

Another straightforward extension is multiple MIVs. In practice, to find a MIV is easier than to find a traditional instrumental variable. It is likely that several MIVs are available. Let $\mathbf{Z}$ be a MIV vector such that for any two realizations $\mathbf{z}_i,\mathbf{z}_j$,

$$E\left(Y_t\left|\mathbf{Z}=\mathbf{z}_i\right.\right)\leq E\left(Y_t\left|\mathbf{Z}=\mathbf{z}_j\right.\right),\forall t,\text{ if }\mathbf{z}_i\leq\mathbf{z}_j,$$

where the meaning of $\mathbf{z}_i\leq\mathbf{z}_j$ is that each component in $\mathbf{z}_i$ is no larger than the corresponding element in $\mathbf{z}_j$.

In the presence of multiple MIVs, Eq. (2), Eq. (3) and Proposition 2 take the same form except that we interpret $1,i,j,n_Z$ as multiple indices in $\max_{1\leq i\leq j}$ and $\min_{j\leq i\leq n_Z}$, etc.

Next, we extend our Bayesian inference procedure to other applied identification problems under the MIV assumption. Kreider and Pepper (2007) consider a partial misreporting problem in which people are surveyed on

their employment and health conditions. However, people may not truthfully report their health, but researchers have some prior information on the truth-telling rate of some subpopulations. In other words, researchers classify the respondents into the verified group and unverified group. Let $L \in \{0,1\}$ be the employment status, $X \in \{0,1\}$ and $W \in \{0,1\}$ be the reported and the true disability status respectively, $Y \in \{0,1\}$ be the verification status, and $Z \in \{z_1, ...z_{n_Z}\}$ be a MIV such that

$$P\left(L=1\,|W, Z=z_i\right) \geq P\left(L=1\,|W, Z=z_j\right), \text{ if } z_i \leq z_j.$$

The joint distribution of $(Z, L, X, Y)$ can be learned from the data, while the joint distribution of $(Z, L, X, Y, W)$ is unknown. For simplicity, we consider an extreme case that the verified group has a 100% truth-telling rate, while the unverified has an accuracy rate $\geq 0\%$ (i.e., no information). Kreider and Pepper (2007) show that the sharp bounds of $P\left(L=1\,|W=1, Z=z_j\right)$ are

$$\begin{aligned} &\max_{z_i \geq z_j} \frac{P\left(L=1, X=1, Y=1\,|Z=z_i\right)}{P\left(X=1, Y=1\,|Z=z_i\right) + P\left(L=0, Y=0\,|Z=z_i\right)} \\ &\leq P\left(L=1\,|W=1, Z=z_j\right) \leq \\ &\min_{z_i \leq z_j} \frac{P\left(L=1, X=1, Y=1\,|Z=z_i\right) + P\left(L=1, Y=0\,|Z=z_i\right)}{P\left(X=1, Y=1\,|Z=z_i\right) + P\left(L=1, Y=0\,|Z=z_i\right)}. \end{aligned} \tag{4}$$

Readers are referred to Proposition 2, corollary 1 in Kreider and Pepper (2007, p.436) for the derivation.

Unlike Eq. (2) or Eq. (3) where the lower bound of interest is the maximum of a linear combination of conditional probabilities, here in Eq. (4) the lower bound is the maximum of a ratio of conditional probabilities. Despite this difference, a similar analytic Bayesian inference can be applied

to the current problem. This is largely due to the flexibility of the Dirichlet distribution in dealing with partition, summation and taking ratios.

Without loss of generality, arrange the values of the MIV as $z_1 \leq ... \leq z_{n_Z}$. We will define a set of symbols close to those in the previous identification problem.

$p_{ijkl} \equiv P\left(Z = z_i, L = j, X = k, Y = l\right)$, $i = 1, ..., n_Z$, $j, k, l = 0, 1$,

$p_{i\cdots} \equiv \sum_{j=0}^{1} \sum_{k=0}^{1} \sum_{l=0}^{1} p_{ijkl}$,

$x_{ijkl} \equiv \frac{p_{ijkl}}{p_{i\cdots}}$, standing for the conditional probability $P\left(L = j, X = k, Y = l \,|Z = z_i\right)$,

$\mathbf{p} \equiv vec\left(\{p_{ijkl}\}_{i=1,j=0,k=0,k=0}^{n_Z,1,1,1}\right)$,

$\mathbf{x}_i \equiv vec\left(\{x_{ijkl}\}_{j=0,k=0,k=0}^{1,1,1}\right)$.

Then Eq. (4) can be written as

$$\max_{1 \leq i \leq j} f_L\left(\mathbf{x}_i\right) \leq P\left(L = 1 \,|W = 1, Z = z_j\right) \leq \min_{j \leq i \leq n_Z} f_U\left(\mathbf{x}_i\right), \tag{5}$$

where

$$f_L\left(\mathbf{x}_i\right) = \frac{x_{i111}}{x_{i111} + x_{i011} + x_{i010} + x_{i000}},$$
$$f_U\left(\mathbf{x}_i\right) = \frac{x_{i111} + x_{i110} + x_{i100}}{x_{i111} + x_{i011} + x_{i110} + x_{i100}}.$$

With a Dirichlet prior on $\mathbf{p}$, we will arrive at the conjugate Dirichlet posterior, say $\mathbf{p} \sim Dir\left[vec\left(\{a_{ijkl}\}_{i=1,j=0,k=0,k=0}^{n_Z,1,1,1}\right)\right]$. Proposition 1 still holds with the independent

$$\mathbf{x}_i \sim Dir\left[vec\left(\{a_{ijkl}\}_{j=0,k=0,k=0}^{1,1,1}\right)\right]$$

Taking ratios of components in $\mathbf{x}_i$, we arrive at a Beta distribution, and then we can derive the posterior distribution of the bounds in Eq. (5).

**Proposition 3.** *The two bounds of* $P(L=1\,|W=1,Z=z_j)$ *defined in Eq. (5), namely* $L \equiv \max_{1\le i\le j} f_L(\mathbf{x}_i)$, $U \equiv \min_{j\le i\le n_Z} f_U(\mathbf{x}_i)$, *has the posterior c.d.f.*

$$F_L(c) = \prod_{1\le i\le j} \frac{B(c; a_{i111}, a_{i011}+a_{i010}+a_{i000})}{B(1; a_{i111}, a_{i011}+a_{i010}+a_{i000})},$$
$$F_U(c) = 1 - \prod_{j\le i\le n_Z} \left[1 - \frac{B(c; a_{i111}+a_{i110}+a_{i100}, a_{i011})}{B(1; a_{i111}+a_{i110}+a_{i100}, a_{i011})}\right],$$

*where* $B(c;a,b)$ *is the incomplete beta function, namely*

$$B(c;a,b) = \int_0^c t^{a-1}(1-t)^{b-1}\,dt.$$

Proof:

Let $L_i = f_L(\mathbf{x}_i)$, $i = 1, ...j$.

For each $i = 1, ..., j$ as given, Let $\{\xi_{ijkl}\}_{j=0,k=0,k=0}^{1,1,1}$ be an array of independently distributed random variables with each component $\xi_{ijkl} \sim Gamma(a_{ijkl}, 1)$.

Let $\eta = \sum_{j=0}^{1}\sum_{k=0}^{1}\sum_{l=0}^{1}\xi_{ijkl}$, and $x_{ijkl} = \frac{\xi_{ijkl}}{\eta}$. It is known that $\mathbf{x}_i \equiv vec\left(\{x_{ijkl}\}_{j=0,k=0,k=0}^{1,1,1}\right)$ has the Dirichlet distribution $Dir\left[vec\left(\{a_{ijkl}\}_{j=0,k=0,k=0}^{1,1,1}\right)\right]$. Then we have

$$L_i = \frac{\xi_{i111}}{\xi_{i111} + (\xi_{i011}+\xi_{i010}+\xi_{i000})},$$

which is recognized as a Beta distribution:

$$L_i \sim Beta(a_{i111}, a_{i011}+a_{i010}+a_{i000}).$$

Proposition 1 implies $\mathbf{x}_1, ..., \mathbf{x}_{n_Z}$ are independent, so are $L_1, ..., L_j$. It follows that $F_L(c) = \prod_{1\le i\le j} F_{L_i}(c)$, where $F_{L_i}(c)$ can be expressed as the ratio of the incomplete and complete beta function.

Similarly, let $U_i = f_U(\mathbf{x}_i)$, $i = j, ..., n_Z$, so we have

$$U_i \sim Beta\left(a_{i111} + a_{i110} + a_{i100}, a_{i011}\right).$$

It follows that $F_U(c) = 1 - \prod_{j \leq i \leq n_Z} [1 - F_{U_i}(c)]$, where $F_{U_i}(c)$ is also a ratio of the incomplete and complete beta function. ■

## 5. An application

As an application to the Bayesian inference on the treatment effect identification with MIVs, we consider the effect of taking extra classes as well as the effect of parents' education on students' academic skills. The data come from National Longitudinal Survey of Youth 1997 (NLSY97). High school students were asked whether they spent any time taking extra classes. Among the 5385 respondents to this question, 1458 provided a positive answer (hereafter refer to them "class-takers", and the rest "non-takers"). Later in 1997-98, most NLSY97 respondents participated the Armed Services Vocational Aptitude Battery (ASVAB), a comprehensive ability test on arithmetic reasoning, word knowledge and general sciences, etc. NLSY97 also collected information on their family background. In our study, we will use their biological father and mother's highest degree as two MIVs, since we believe the expectation of conjectured class-takers' test scores (and non-takers' test scores as well) vary monotonically with fathers' (mothers') education levels, with mothers' (fathers') education being the same. To increase the identification power, the MTR assumption is also imposed, which states that ceteris paribus, everyone's conjectured class-taker's test scores is higher than his or her conjectured non-taker's test scores. To put it simply, class taking is good

for everyone's score. The MIV and MTR assumptions represent two distinct ways to improve test scores. One is an internal source, which guarantees a higher score as long as one endeavors to take extra classes. The other one is an external source, which asserts knowledgeable parents' guidance will help children in the sense that average score increases, but not necessarily for everyone. Therefore, for the distribution of conjectured response variables, both the conditional and unconditional mean response are of interest when we study the sources conducive to academic performance.

Our inference approach requires discretization of all variables. The treatment variable ($D$), namely taking extra classes, is binary. The two MIVs ($\mathbf{Z}$) take on one of the 7 values: none degree, GED, high school diploma, associate/Junior college, Bachelor's degree, Master's degree, Ph.D. or professional degrees. As for the observed response variable ($Y$), we discretize it into 11 evenly spaced grids $(0, 10, 20, ..., 100)$, rounding the original ASVAB math-verbal score percent to the nearest grid. Clearly, the lower bound of the conjectured response variables ($Y_0$, $Y_1$) is 0, and the upper bound is 100.

The descriptive statistics of variables are provided in Table 1. The largest two cohorts are students with both parents being high school graduates (31.3%), and students with both parents holding Bachelor's degrees (5.9%). So we mainly compare distributions conditional on those two groups. On the basis of Proposition 2, the analytic c.d.f. of lower and upper bounds of $E(Y_1 | Z)$, $E(Y_0 | Z)$ are calculated with a uniform prior (setting all Dirichlet hyperparameters equal to one). The c.d.f. is evaluated at 500 points along the interval $[0, 100]$ and by differencing we have density estimates. Figure 1 displays the distribution of bounds of expected $Y_1$ and $Y_0$ conditional on par-

ents' education, and Table 2 presents corresponding summary statistics on those posterior distributions. First we compare the left-top and left-bottom panels of Figure 1, which contrasts the role of educated parents on children's academic performance. Parents' education substantially shifts the average conjectured class-takers' test scores. The mean lower bound of expected $Y_1$ is 46.5 with the standard deviation 2.3 conditional on high school graduated parents, while the mean lower bound conditional on parents with Bachelor's degrees is 66.2 with the standard deviation 1.7. For children who do not take extra classes, parents' education also have substantial impact on children's test scores, which can be seen from the right-top and right-bottom two panels. Next, a horizontal comparison of the top two (and the bottom two) panels of Figure 1 reveals the effects of extra-class taking on conjectured response variables, conditional on the same parents' education level. Even if we compare the mean upper bound of the expected $Y_0$ with the mean lower bound of the expected $Y_1$, we still find an improvement of ASVAB score by 17.3 percentage points when parents are college graduates, and by 3.3 percentage points when parents are high school graduates.

By marginalization in accordance with the distribution of the MIVs, we find the distribution of unconditional mean of $Y_1$, $Y_0$ as well as the ATE. Since analytic distributions are not available, 50000 draws from the posterior Dirichlet distribution are generated to simulated the posterior bounds of expected $Y_1$, $Y_0$ and $Y_1 - Y_0$. The relevant distributions are graphed in Figure 2, 3 and summarized in Table 3. The lower bound of $E(Y_1)$ has the mean 51.7 and the upper bound of $E(Y_0)$ averages 43.3. The ATE is positive for the sure, since the lower bound of $E(Y_1 - Y_0)$ is distributed with the mean

8.5 and the 95% HPD credible interval is $(4.8, 12.7)$. The mean of the upper bound of $E(Y_1 - Y_0)$ is 33.9, with the 95% HPD interval $(26.6, 40.5)$. This seems to suggest attending some extra classes is worthwhile.

There is a note to this application. One might argue that taking extra classes may not be beneficial for everyone. This concern is legitimate. However, if we give up the MTR assumption, the MIV assumption alone will not yield much identification power. Note that students who received the treatment accounts for 29%, while the rest 71% did not attend extra classes. Put aside sampling variations and let us have a quick account of the lower bounds of $E(Y_1)$ and $E(Y_0)$. $E(Y_1 | D = 0)$, $E(Y_0 | D = 1)$ are unobservable and have to be assigned 0 in the absence of the MTR assumption. Then we compute the lower bounds of $E(Y_1)$ as $E(Y_1 | D = 1) \times 21\% + 0 \times 79\%$ and that of $E(Y_0)$ as $E(Y_0 | D = 0) \times 79\% + 0 \times 21\%$. Even if $Y_1 > Y_0$ everywhere, the large weights on zeros may cause our estimated lower bound of $E(Y_1)$ lower than that of $E(Y_0)$, and lower than the upper bound of $E(Y_0)$ as well. Therefore the lower bound of the ATE could be negative. The mechanism of MIV identification is to divide the population into sub-populations and repeat the above calculation in each sub-population. By taking the maximum and minimum, the MIV assumption pulls up the lower bound of $E(Y_1)$ and pushes down the upper bound of $E(Y_0)$, but it still has difficulty reversing the sign of the ATE when most respondents do not receive the treatment. However, if one adopts the MTR assumption, the lower bounds of $E(Y_1)$ can be identified as $E(Y_1 | D = 1) \times 21\% + E(Y_0 | D = 0) \times 79\%$, which is a substantial improvement. In that sense, we argue that if the non-treated constitute the majority in the sample, the MIV assumption alone is not very

likely to have significant identification power.

Chen, M., Shao, Q., 1998. Monte carlo estimation of bayesian credible and hpd intervals. Journal of Computational and Graphical Statistics 8, 69–92.

Chernozhukov, V., Hong, H., Tamer, E., 2007. Estimation and confidence regions for parameter sets in econometric models. Econometrica 75 (5), 1243–1284.

Chernozhukov, V., Lee, S. S., Rosen, A., 2009. Intersection bounds: estimation and inference. CeMMAP working papers CWP19/09.

David, H. A., 1981. Order Statistics. Wiley.

Gil-Pelaez, J., 1951. Note on the inversion theorem. Biometrika 38 (3-4), 481–482.

Imbens, G. W., Manski, C. F., 2004. Confidence intervals for partially identified parameters. Econometrica 72 (6), 1845–1857.

Imhof, J. P., 1961. Computing the distribution of quadratic forms in normal variables. Biometrika 48 (3-4), 419–426.

Kreider, B., Pepper, J. V., 2007. Disability and employment: Reevaluating the evidence in light of reporting errors. Journal of the American Statistical Association 102, 432–441.

Manski, C. F., 1997. Monotone treatment response. Econometrica 65 (6), 1311–1334.

Manski, C. F., Pepper, J. V., 2000. Monotone instrumental variables, with an application to the returns to schooling. Econometrica 68 (4), 997–1012.

Manski, C. F., Pepper, J. V., 2009. More on monotone instrumental variables. Econometrics Journal 12 (s1), S200–S216.

Moschopoulos, P., 1985. The distribution of the sum of independent gamma random variables. Annals of the Institute of Statistical Mathematics 37, 541–544.

Provost, S. B., Cheong, Y.-H., 2000. On the distribution of linear combinations of the components of a dirichlet random vector. Canadian Journal of Statistics 28 (2), 417–425.

Qian, H., 2010. Sampling variation and monotone instrument variable under discrete distributions (manuscript).

Rosen, A. M., 2008. Confidence sets for partially identified parameters that satisfy a finite number of moment inequalities. Journal of Econometrics 146 (1), 107–117.

Ross, A., 2010. Computing bounds on the expected maximum of correlated normal variables. Methodology and Computing in Applied Probability 12, 111–138.

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Extra class taking (Treatment variable) | | | | | | | | | | | |
| Value | 0 | 1 | | | | | | | | | |
| Percent | 71 | 29 | | | | | | | | | |
| ASVAB score (Response variable) | | | | | | | | | | | |
| Value | 0 | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 |
| Percent | 5 | 11 | 11 | 10 | 10 | 10 | 9 | 10 | 9 | 9 | 5 |
| Mother highest degree (MIV 1) | | | | | | | | | | | |
| Value | 1 | 2 | 3 | 4 | 5 | 6 | 7 | | | | |
| Percent | 14 | 5 | 47 | 11 | 16 | 6 | 1 | | | | |
| Father highest degree (MIV 2) | | | | | | | | | | | |
| Value | 1 | 2 | 3 | 4 | 5 | 6 | 7 | | | | |
| Percent | 16 | 4 | 48 | 8 | 15 | 6 | 3 | | | | |

Table 1: Descriptive statistics on the treatment variable, response variable, and MIVs.

| | Mean | Std | Median | Mode | 95% HPD |
|---|---|---|---|---|---|
| Distribution of lower bounds | | | | | |
| $E(Y_1 \mid Z = Bachelor)$ | 66.19 | 1.66 | 66.13 | 66.13 | [62.93 , 69.14] |
| $E(Y_0 \mid Z = Bachelor)$ | 41.11 | 1.86 | 40.88 | 40.28 | [36.67 , 46.09] |
| $E(Y_1 \mid Z = HighSchool)$ | 46.46 | 2.25 | 45.89 | 44.69 | [42.89 , 50.70] |
| $E(Y_0 \mid Z = HighSchool)$ | 33.92 | 1.76 | 33.47 | 31.86 | [29.86 , 39.08] |
| Distribution of upper bounds | | | | | |
| $E(Y_1 \mid Z = Bachelor)$ | 71.43 | 5.12 | 71.94 | 72.95 | [61.32 , 80.56] |
| $E(Y_0 \mid Z = Bachelor)$ | 48.87 | 4.83 | 49.10 | 49.50 | [39.28 , 57.72] |
| $E(Y_1 \mid Z = HighSchool)$ | 68.66 | 4.25 | 69.14 | 70.14 | [60.52 , 76.15] |
| $E(Y_0 \mid Z = HighSchool)$ | 43.18 | 1.95 | 43.69 | 43.89 | [38.88 , 45.69] |

Table 2: Summary statistics on the distribution of lower and upper bounds for the expectation of the conjectured ASVAB score $(Y_1, Y_0)$ conditional on parents education levels.

| | Mean | Std | Median | Mode | 95% HPD |
|---|---|---|---|---|---|
| Distribution of lower bounds | | | | | |
| $E(Y_1)$ | 51.72 | 0.99 | 51.63 | 51.50 | [49.92 , 53.69] |
| $E(Y_0)$ | 34.74 | 1.38 | 34.56 | 34.27 | [32.34 , 37.60] |
| $E(Y_1) - E(Y_0)$ | 8.46 | 2.08 | 8.19 | 7.82 | [4.78 , 12.72] |
| Distribution of upper bounds | | | | | |
| $E(Y_1)$ | 68.69 | 3.37 | 69.10 | 69.74 | [62.03 , 74.84] |
| $E(Y_0)$ | 43.26 | 1.86 | 43.56 | 44.09 | [39.39 , 46.38] |
| $E(Y_1) - E(Y_0)$ | 33.95 | 3.64 | 34.33 | 35.07 | [26.58 , 40.53] |

Table 3: Summary statistics on the distribution of lower and upper bounds for the unconditional expectation of the conjectured ASVAB score $(Y_1, Y_0)$.

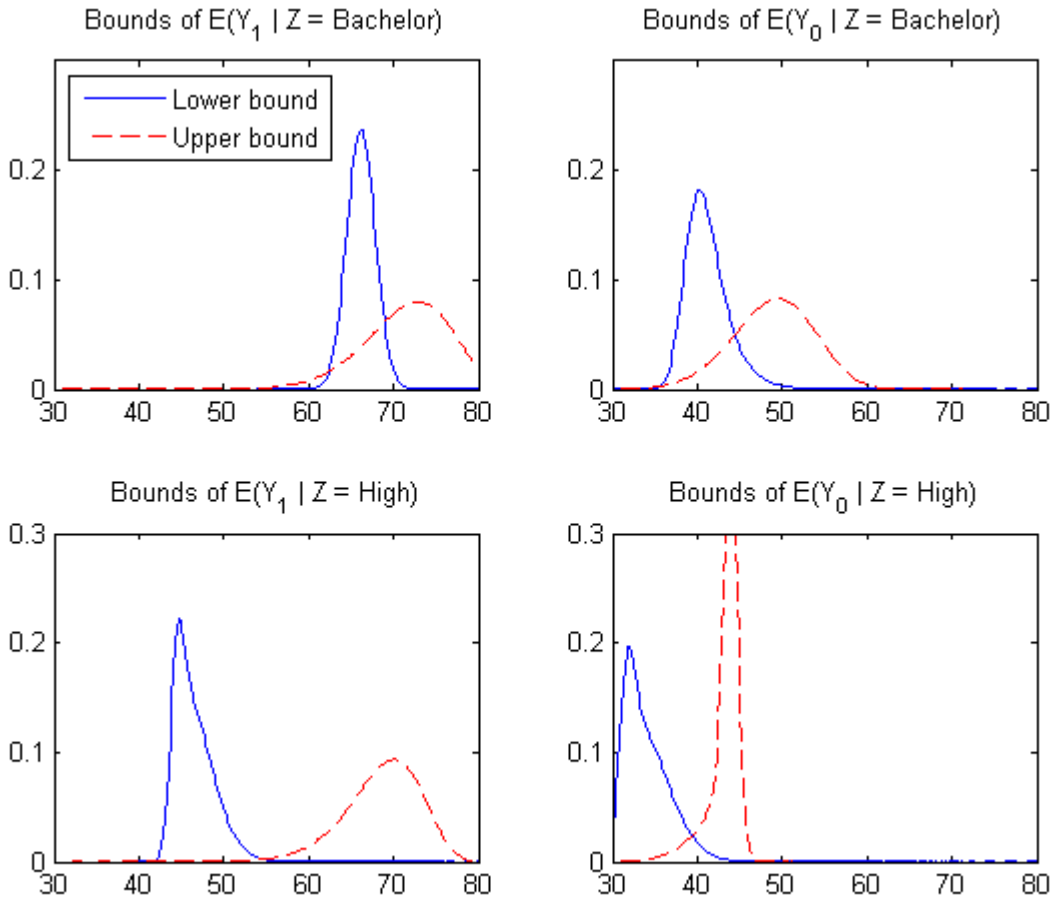


Figure 1: Analytic lower and upper bounds for the expectation of the conjectured ASVAB score $(Y_1, Y_0)$ conditional on parents education levels. Density estimates of the score distribution obtained from the analytic c.d.f. are plotted in four graphs. Vertical comparison of two graphs shows the effect of parents education, while horizontal comparison of two graphs shows the effect of taking extra classes.

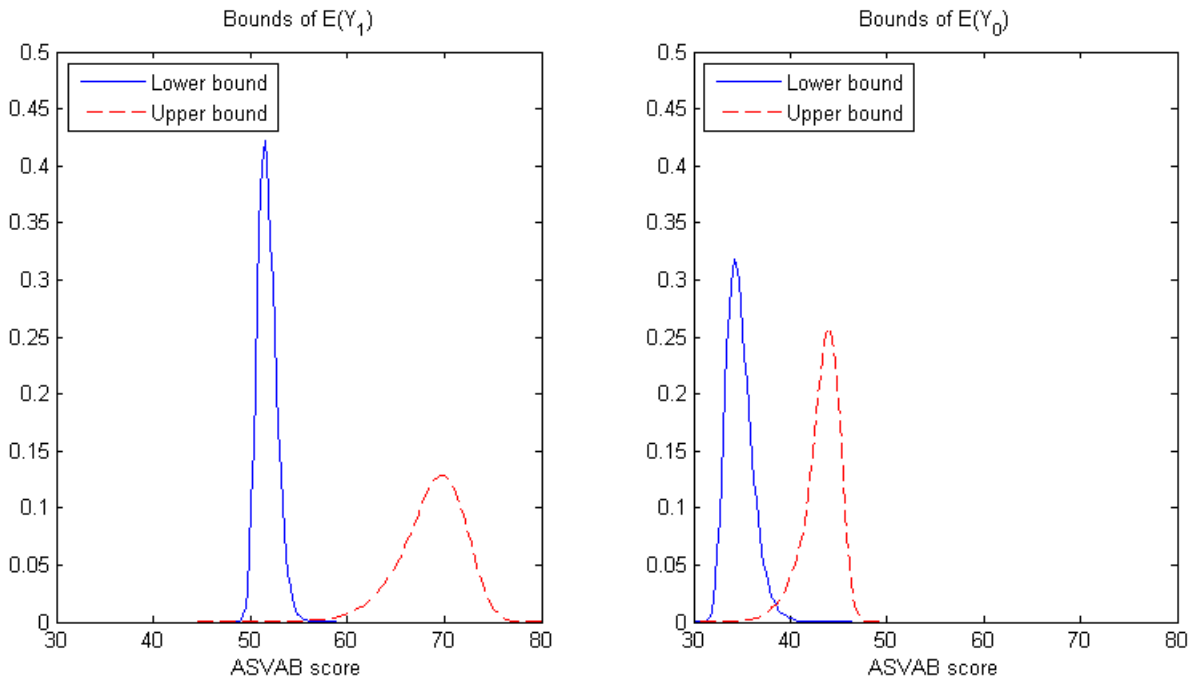


Figure 2: Simulated lower and upper bounds for the unconditional expectation of conjectured ASVAB score $(Y_1, Y_0)$. Kernel density estimates are obtained from 50000 posterior draws.

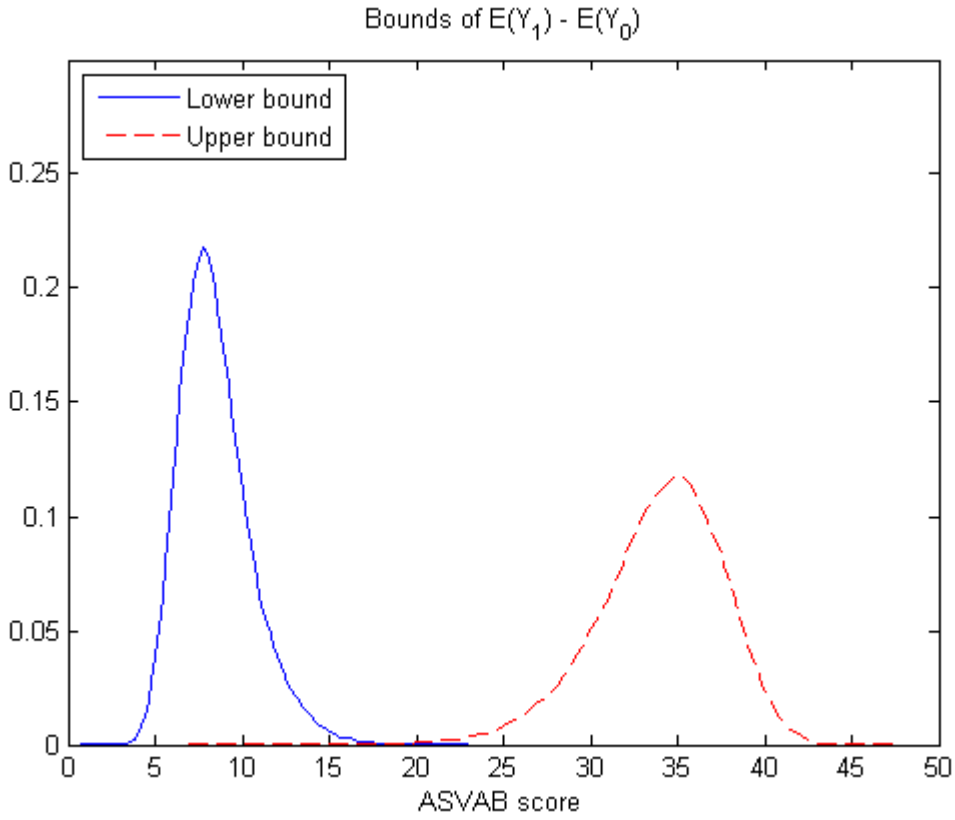


Figure 3: Simulated lower and upper bounds for the average treatment effect. Kernel density estimates are obtained from 50000 posterior draws.